Representation Engineering: A Top-Down Approach to AI Transparency

Year: 2023 | Citations: 665 | Authors: Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo

---

Abstract

In this paper, we identify and characterize the emerging area of representation engineering (RepE), an approach to enhancing the transparency of AI systems that draws on insights from cognitive neuroscience. RepE places population-level representations, rather than neurons or circuits, at the center of analysis, equipping us with novel methods for monitoring and manipulating high-level cognitive phenomena in deep neural networks (DNNs). We provide baselines and an initial analysis of RepE techniques, showing that they offer simple yet effective solutions for improving our understanding and control of large language models. We showcase how these methods can provide traction on a wide range of safety-relevant problems, including honesty, harmlessness, power-seeking, and more, demonstrating the promise of top-down transparency research. We hope that this work catalyzes further exploration of RepE and fosters advancements in the transparency and safety of AI systems.